Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Hoshida Y, Villanueva A, Kobayashi M, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. N Engl J Med 2008;359:1995-2004. DOI: 10.1056/NEJMoa0804525.

RNA extraction

Tumor and adjacent liver tissues were macro-dissected from 10 micron formalin-fixed, paraffin-embedded (FFPE) tissue sections. Absence of microvascular tumor invasion in the adjacent liver tissue was confirmed using H & E staining of consecutive sections. Using 3-4 sections for each sample, total RNA was extracted using the High Pure RNA Paraffin kit (Roche) as directed by the manufacturer (training set) or TRIzol LS reagent (Invitrogen) in a semi-automated 96-well plate format based on the manufacturer's instructions (validation set).

Gene expression arrays for FFPE tissues

DASL assay

To profile randomly fragmented mRNA extracted from FFPE tissue (FFPE-RNA), we employed the cDNA-mediated Annealing, Selection, extension and Ligation (DASL) assay (Illumina)[1, 2]. Briefly, fragmented FFPE-RNA is converted into cDNA using random primers. For each target site on the cDNA, a pair of query oligos separated by a single nucleotide is annealed to the cDNA, and, the gap between the query oligos is extended and ligated to generate a PCR template. A pair of universal PCR primers is then used for amplification, and linearly amplified PCR products are hybridized to a bead microarray. The array was then scanned by a BeadArray Reader (Illumina).

Number of microarray probes assigned to each gene

Missing signals due to RNA degradation is one of the major concerns in profiling FFPE tissues. For this reason, a commercially available panel of 502 cancer-related genes for DASL assay (Cancer Panel, Illumina) assigns 3 independent probes to each gene, with the expectation that this would maximize data quality. However, the use of multiple probes per gene diminishes the number of transcripts that can be assayed per array (given a fixed number of probes per array). We therefore sought to experimentally determine the effect of reducing the number of probes per gene, so as to facilitate covering a larger number of genes with the same total number of probes. We randomly picked a single probe from among the 3 probes assigned to each gene, and evaluated how the single probe dataset performed in sample clustering and marker gene selection analyses.

First, by picking a single probe for each gene, 5~7% of measurements fell below the

level of negative control probes, suggesting either missing signals due to RNA degradation or suboptimal probe sequence (**Supplementary Figure 1A**). However, we found that such probe drop-out had little effect on overall performance of the arrays. For example, a prostate cancer vs. normal distinction was not affected by the single probe picking (**Supplementary Figure 1B**). This suggests that profiling 100s~1000s genes can compensate for the slight increase in noise caused by RNA degradation. In marker gene analysis, only a small number of genes were dropped from the top marker gene list (indicating a small number of false negatives), and no genes came to the top of the marker list in the single probe data but were absent in the dataset using 3 probes per gene (indicating no false positives) (**Supplementary Figure 1C**).

Designing a 6,000-gene DASL assay

We sought to identify ~ 6,000 maximally informative transcripts that could be used for genome-wide discovery on the DASL platform (configured as 4 x 1536 assays utilizing one probe per gene). To address this, we analyzed a large collection of Affymextrix transcriptome datasets profiling cancer and normal tissues.[3, 4] This analysis revealed that the expression signals from ~ one third of the genes on most genome-wide arrays were "absent" (**Supplementary Figure 8**). This suggests that a substantial proportion of the genome is infrequently expressed, and therefore might be omitted without great consequence. By excluding such genes, we aimed to define a generic minimum subset of genome representing the global structure of the entire transcriptome.

We designed a set of query oligos (i.e., probes) to profile transcriptionally informative genes that might be useful for signature discovery and validation. To this end, we selected genes with the largest variation across samples in a large collection of previously generated Affymetrix microarray datasets spanning 24 studies, 2,149 samples, and 15 tissue types (**Supplementary Table 8**). After filtering out genes with less than a 3-fold difference and less than 100 units between the maximum and minimum signals across the dataset, the coefficient of variation (CV) was calculated and summarized onto the NCBI's RefSeq gene IDs to compute a priority score for each gene, and genes were rank-ordered according to this score (**Supplementary Figure 9A**). An examination of published marker genes from recent studies indicated

that our list of 6,000 genes represented 70-90% of these genes, indicating that the 6,000 gene array was more informative than a random collection of 6,000 genes (which might be expected to capture only ~ 25% of reported markers) (**Supplementary Figure 9B**). We then designed query oligos for the top informative 6,100 genes (NCBI's Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/, platform ID GPL5474).

Quality assessment of DASL profile

As a quality measure of the DASL gene expression profile, we calculated the proportion of gene probes with a "present" signal (%P-call), which is expected to be similar across samples of a given tissue type (e.g. HCC). (Of note, the "present" call rate drops precipitously when degraded RNA typical of FFPE tissues is analyzed on conventional microarrays such as Affymetrix arrays). The "present" call was computed based on built-in negative control probes (GenePattern, IlluminaDASL pipeline). In a pilot experiment performed on 10 prostate cancer tissues, we observed %P-call of ~ 75% in 2 samples fixed 24 years before RNA extraction, which was comparable to a sample fixed 7 years ago (**Supplementary Figure 9C**), indicating that data quality is not directly correlated with age of the sample.

Poor quality profiles were detected and removed as follows. We set a "median" array as a representative sample in a dataset by calculating the median for each gene. The poor quality, outlier profiles were defined based on dissimilarity to the "median" array measured by Pearson correlation coefficient. In the plot of the correlation to %P-call, we observed that the correlation sharply started to drop as %P-call became smaller than a certain value. This likely indicates that the samples with %P-call smaller than this value have severe RNA degradation affecting sensitivity of gene expression signal detection. Based on this plot, we set a quality threshold of %P-call for each tissue type to assure a minimum correlation coefficient of 0.7 for the majority of the samples (we set the %P-call quality thresholds of 65% and 70% for tumor and adjacent liver tissues, respectively, **Supplementary Figure 10**). Failure of the profiling, i.e., %P-call less than 70% in adjacent liver set, was not associated with clinical variables including age ($p=0.49$), sex ($p=0.78$), existence of cirrhosis ($p=1.00$), Child-Pugh stage ($p=0.11$), HCC etiology ($p>0.70$), or age of the FFPE block (>10 years, $p=0.30$).

The same %P-call threshold was applied for the validation set. After eliminating samples with poor quality data, the raw data were normalized using the cubic spline algorithm[5] using the IlluminaDASL pipeline within GenePattern. Only gene probes with a minimal 3-fold differential expression and absolute difference >500 units across the samples were included after applying floor and ceiling values of 200 and 80,000 units, respectively.

Comparison of gene expression profiles between intact and FFPE-RNA

First, we evaluated the extent of correlation of gene expression profile of FFPE tissue with that of fresh tissue at the level of individual genes. To ensure a uniform population of cells being subjected to the fresh and fixed analysis, we used cell lines (as opposed to tissues, which have greater intra-tissue variability which would become a confounding factor in these analyses). DHL4 and Hela cell lines were cultured, harvested, and split into two halves. Total RNA was immediately extracted from one half, and the other half was fixed with formalin and embedded in a paraffin block. Total RNA was also extracted from the FFPE block using the protocol described in the **Methods** section. All RNA samples were profiled using the DASL assay, and fold changes were calculated for each gene in a comparison between DHL4 and Hela cell lines. The plot of the fold changes for the intact and FFPE cell lines showed moderate correlation (Pearson correlation coefficient 0.61, $p<0.001$, **Supplementary Figure 11**). At the higher fold changes in the fresh RNA profiles, the vast majority of the genes showed concordant gene expression changes in the FFPE profiles (**Supplementary Table 9**).

Next, we determined whether the DASL profile of FFPE tissue recapitulates the biologically relevant information observed in the profile of fresh frozen tissue. For this analysis, we turned to prostate cancer, for which there exists an abundance of published microarray data derived from frozen tumor and normal tissues. We identified 200 marker genes that reflect the tumor vs. normal prostate distinction based on a meta-analysis of 7 published frozen sample-based microarray datasets collected in a cancer transcriptome database (Oncomine, http://www.oncomine.org). Among those genes, 180 genes (90%) are included in our 6,100 informative gene panel. Based on the expression pattern of those marker genes, we classified a

collection of FFPE tumor and normal prostate samples using a nearest template prediction method (see Data analysis section). We observed 100% accurate prediction with statistical significance (false discovery rate <0.05, **Supplementary Figure 12**), indicating that the 6,000-gene DASL assay robustly identifies biologically meaningful patterns in FFPE tissues. We also performed a meta-analysis of 3 independent frozen sample-based HCC datasets including 232 samples to define common subclasses of HCC, and found that the molecular subclasses identified in the frozen tissues were also seen in the profiles of 118 FFPE HCC tissues profiled by DASL (manuscript in preparation). We therefore conclude that our 6,000-gene DASL assay accurately recapitulates the gene expression profile of fresh frozen tissues in archived, FFPE material.

Data availability

Microarray datasets are available through Gene Expression Omnibus (GSE10143) or our web site at http://www.broad.mit.edu/cancer/pub/HCC.

Data analysis

Definition of clinical outcome

While HCC is the cause of death in most patients with the disease, some patients die of liver failure or other causes attributable to cirrhosis in the absence of progressive HCC (7 of the 39 deaths in our study died of non-HCC causes). Accordingly, we chose HCC-related mortality (disease-specific death) as the principal clinical endpoint for the survival-predictive signature discovery, defined as follows: (1) tumor occupying more than 80% of the liver, (2) portal venous tumor thrombus (PVTT) proximal to the second bifurcation, (3) obstructive jaundice due to tumor, (4) distant metastasis, or (5) variceal hemorrhage with PVTT proximal to the first bifurcation. The commonly used definition of "late recurrence" was tumor recurrence appearing more than 2 years after surgery[6, 7]. For late recurrence prediction, early recurrences were treated as censored observations.

Prognostic prediction

Most outcome prediction studies discretize outcome in a binary fashion, creating two classes of patients: those with good outcome, and those with bad outcome. Unfortunately, this approach requires creating a boundary between the two groups that

is often not obvious, and the approach works poorly with patients of intermediate outcome. In this study, we used non-discretized, censored survival time to select signature genes in order to not sacrifice sample size and to avoid the problem of setting an arbitrary cut-off of survival time. In addition, we sought to determine whether the expression of poor- and good-prognosis signature genes were coordinately regulated in a given sample. That is, it was expected that the poor signature genes would be ON (or up) and the good signature genes would be OFF (or down) in a "poor" survival sample. To evaluate this, we designed a simple nearest neighbor-based method assessing a sample's proximity to a hypothetical representative sample (template) of poor or good survival. This approach allowed us to perform single sample-based outcome prediction. The details of the method are described below.

Genes positively or negatively correlated with HCC-related survival or time-to-recurrence were selected using the Cox score[8, 9] using the following formula.

$$d = \left[\sum_{k=1}^{K} (x_k^* - d_k \bar{x}_k)\right] \Big/ \left[\sum_{k=1}^{K} (d_k / m_k) \sum_{i \in R_k} (x_i - \bar{x}_k)^2\right]^{1/2}$$

where i is indices of samples, x_i is gene expression level for sample i, t_i is time for sample i, $k \in 1, \cdots, K$ is indices of unique death times $z_1, z_2, \ldots, z_K$, d_k is number of deaths at time z_k, m_k is number of samples in $R_k = i : t_i \geq z_k$, $x_k^* = \sum_{t_i = z_k} x_i$, and $\bar{x}_k = \sum_{i \in R_k} x_i / m_k$. Prediction analysis was performed by evaluating the expression status of the signature using the nearest template prediction (NTP) method as implemented in the NearestTemplatePrediction module of the GenePattern analysis toolkit. Briefly, a hypothetical sample serving as the template of "poor" outcome was defined as a vector having the same length as the predictive signature. In this template, a value of 1 was assigned to "poor" outcome-correlated genes and a value of -1 was assigned to "good" outcome-correlated genes, and then each gene was weighted by the absolute value of the corresponding Cox score. The template of "good" outcome was similarly defined. For each sample, a prediction was made based on the proximity measured by the cosine distance to either of the two templates. Significance for the proximity was estimated by comparison to a null distribution generated by randomly picking (1,000 times) the same number of marker genes from

the microarray data for each sample, and correcting for multiple hypothesis testing using the false discovery rate (FDR)[10]. A sample closer to the template of "poor" outcome with an FDR <0.05 was predicted as having poor outcome.

Study design to define outcome-predictive signature

Tumor and adjacent non-tumor liver tissues from the training set were profiled separately to define an outcome-predictive signature (**Figure 1**). The signature was first internally validated in the training set using a leave-one-out cross-validation prediction procedure. A single sample was left out one-by-one and an outcome-correlated signature was selected from the remaining samples (selecting marker genes based on permutation test p-value less than 0.05). A predicted label was assigned to the left-out sample based on the closest "template" using NTP algorithm. Only genes selected in each of the leave-one-out trials were included in the outcome-predictive signatures tested on the validation set.

Gene Set Enrichment Analysis

Functional annotation of the survival signature was performed by Gene Set Enrichment Analysis (GSEA)[11]. We evaluated two categories of annotated gene sets: target genes of experimental perturbation (473 sets) and literature-based curated pathway gene sets (150 sets) collected in our molecular signature database (MSigDB, http://www.broad.mit.edu/gsea/msigdb/index.jsp).

Survival data analysis

Survival difference was evaluated by the log-rank test, and survival association of clinical variables and the signatures was assessed by Cox regression analysis (Survival Analysis modules, GenePattern). First, we evaluated well-accepted clinical predictors of HCC outcome[6, 12]: AFP, multinodularity, and vascular invasion, by univariate analysis. Only variables with statistical significance ($p<0.05$) were further evaluated by multivariate analysis. The hazard rate for tumor recurrence was calculated as previously described[7, 13] to estimate the pattern of HCC recurrence over time after surgery. GenePattern modules and pipeline used in this study are available from http://www.broad.mit.edu/cancer/software/genepattern/. All other clinical data analyses were performed using the R statistical package (http://www.r-project.org).

Clonality analysis

We profiled 5 pairs of primary and recurrent HCC tumors, 2 pairs of adjacent non-tumor liver tissues, and Hela cells for SNPs using the LinkagePanel beadarray (Illumina) according to the manufacturer's instructions[14]. Genotype calls were generated using BeadStudio software (Illumina). In order to address whether primary tumors and recurrences likely derived from the same clone, we analyzed the pattern of heterozygosity in each of the samples. In particular, we counted how many loci appeared homozygous in the primary tumor, yet were called as heterozygous at recurrence. Such cases would suggest that primaries and recurrences derived from different clones, given that regions of LOH in a primary tumor (appearing homozygous on SNP arrays) would likely appear the same in recurrences if the recurrences derived from the same clone (**Supplementary Table 7A**). We similarly analyzed pairs of primary and recurrence/metastasis tumor tissues in endometrial (n=3), ovarian (n=4), lymphoma (n=6) and renal (n=3) cancers to estimate the same measure of clonality in other, non-HCC tumor types (**Supplementary Table 7B**). Strikingly, the HCC pairs showed a significantly higher proportion of loci that appeared homozygous in the primary tumor, yet appeared heterozygous at recurrence (p=0.008, Wilcoxon rank sum test). Similarly, there were more loci that were heterozygous in the HCC primary and homozygous at recurrence, compared to other tumor types (p=0.001) (**Supplementary Figure 7**).

Outcome prediction using HCC tissue data

We determined whether other machine-learning classifiers based on the binary classes (i.e., "good" and "poor" prognosis) predict outcome in the profiles of HCC tissues. We tested multiple classification methods including Classification of Regression Tree (CART), k-nearest neighbor (k-NN), weighted voting (WV), and support vector machine (SVM), but as shown in **Supplementary Table 10**, these methods also failed to yield statistically significant predictions (p=0.34 for survival and p=0.92 for recurrence. Log-rank test). This result indicates that the failed HCC tissue-based outcome prediction by our method is not due to selection of classification algorithm.

Survival signature in fresh frozen non-tumor liver

We confirmed that the survival signature was readily detectable in a publicly available, independent dataset of fresh frozen non-tumor liver tissues (GSE6764)

(**Supplementary Figure 13**). Prediction was performed using the nearest template prediction method (see description in **Supplementary Appendix**).

Patient survival in validation set according to geographic site

A trend toward survival separation was also seen within each geographic site in the validation set (i.e., U.S., Spain and Italy), although this did not reach statistical significance due to the small sample size and/or insufficient follow-up time in each site (**Supplementary Figure 14**).

References

1. Fan JB, Yeakley JM, Bibikova M, et al. A versatile assay for high-throughput gene expression profiling on universal array matrices. Genome Res 2004;14(5):878-85.
2. Bibikova M, Talantov D, Chudin E, et al. Quantitative gene expression profiling in formalin-fixed, paraffin-embedded tissues using universal bead arrays. Am J Pathol 2004;165(5):1799-807.
3. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001;98(26):15149-54.
4. Su AI, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 2004;101(16):6062-7.
5. Workman C, Jensen LJ, Jarmer H, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol 2002;3(9):research0048.
6. Bruix J, Sherman M. Management of hepatocellular carcinoma. Hepatology 2005;42(5):1208-36.
7. Imamura H, Matsuyama Y, Tanaka E, et al. Risk factors contributing to early and late phase intrahepatic recurrence of hepatocellular carcinoma after hepatectomy. J Hepatol 2003;38(2):200-7.
8. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004;2(4):E108.
9. Significance Analysis of Microarrays: Users manual, http://www-stat.stanford.edu/~tibs/SAM/sam.pdf.
10. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19(3):368-75.
11. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545-50.
12. Llovet JM, Burroughs A, Bruix J. Hepatocellular carcinoma. Lancet 2003;362(9399):1907-17.
13. Mazzaferro V, Romito R, Schiavo M, et al. Prevention of hepatocellular carcinoma recurrence with alpha-interferon after liver resection in HCV cirrhosis. Hepatology 2006;44(6):1543-54.
14. Lips EH, Dierssen JW, van Eijk R, et al. Reliable high-throughput genotyping and loss-of-heterozygosity detection in formalin-fixed, paraffin-embedded tumors using single nucleotide polymorphism arrays. Cancer Res 2005;65(22):10188-91.

Supplementary Table 1
Univariate Cox regression of clinical variables for patient survival (Training set)

Variable	Category	Hazard ratio	95% confidence interval		p-value
			low	high	
Age	≥ 60	0.75	0.37	1.49	0.40
Sex (male)	male	0.41	0.14	1.16	0.09
HBV		0.61	0.23	1.57	0.30
HCV		2.18	0.84	5.69	0.11
Alcohol		2.60	0.79	8.59	0.12
BCLC stage	B (vs. 0/A)	1.45	0.44	4.77	0.54
	A/B (vs. 0)	1.90	0.86	4.20	0.11
Tumor diameter (cm)	≥ 3cm	1.21	0.60	2.43	0.60
Tumor differentiation	Moderate (vs. Well	0.84	0.39	1.85	0.67
	Poor (vs. Well)	0.67	0.23	2.02	0.48
Vascular invasion		2.11	0.50	8.94	0.31
Cirrhosis		1.90	0.78	4.58	0.16
AFP (ng/mL)	≥ 100	0.95	0.47	1.94	0.89
Platelet count ($\times 10^9$/L)	< 10.0	1.68	0.86	3.28	0.13

HB: hepatitis B, HCV: hepatitis C virus, AFP: alpha-fetoprotein

Supplementary Table 2
Survival signature genes defined in adjacent liver tissue (defined in Training set)

Genes correlated with poor survival

Probe ID	GeneID	Gene symbol	Description	Cox score
DAP1_5052	2488	FSHB	follicle stimulating hormone, beta polypeptide	4.80
DAP1_0153	6456	SH3GL2	SH3-domain GRB2-like 2	4.21
DAP1_2390	23029	RBM34	RNA binding motif protein 34	4.19
DAP3_3833	23397	NCAPH	non-SMC condensin I complex, subunit H	4.02
DAP1_0623	1950	EGF	epidermal growth factor (beta-urogastrone)	3.97
DAP1_5926	7204	TRIO	triple functional domain (PTPRF interacting)	3.90
DAP3_3842	1293	COL6A3	collagen, type VI, alpha 3	3.87
DAP1_0171	3983	ABLIM1	actin binding LIM protein 1	3.86
DAP3_0607	3680	ITGA9	integrin, alpha 9	3.81
DAP4_5449	4922	NTS	neurotensin	3.78
DAP3_1324	5055	SERPINB2	serpin peptidase inhibitor, clade B (ovalbumin), member 2	3.69
DAP3_1228	4316	MMP7	matrix metallopeptidase 7 (matrilysin, uterine)	3.59
DAP3_4010	5593	PRKG2	protein kinase, cGMP-dependent, type II	3.44
DAP4_1888	9170	EDG4	endothelial differentiation, lysophosphatidic acid G-protein-coupled	3.40
DAP3_0208	4843	NOS2A	nitric oxide synthase 2A (inducible, hepatocytes)	3.33
DAP1_4004	2043	EPHA4	EPH receptor A4	3.25
DAP4_2216	6672	SP100	SP100 nuclear antigen	3.19
DAP2_0010	2326	FMO1	flavin containing monooxygenase 1	3.04
DAP3_2729	2877	GPX2	glutathione peroxidase 2 (gastrointestinal)	3.02
DAP3_5508	496	ATP4B	ATPase, H+/K+ exchanging, beta polypeptide	2.99
DAP1_5176	8870	IER3	immediate early response 3	2.98
DAP4_5988	7456	WIPF1	WAS/WASL interacting protein family, member 1	2.98
DAP1_3877	3489	IGFBP6	insulin-like growth factor binding protein 6	2.93
DAP1_0897	1501	CTNND2	catenin (cadherin-associated protein), delta 2 (neural plakophilin-related arm-repeat protein)	2.92
DAP3_5371	2200	FBN1	fibrillin 1	2.91
DAP4_5022	2629	GBA	glucosidase, beta; acid (includes glucosylceramidase)	2.85
DAP1_4874	22858	ICK	intestinal cell (MAK-like) kinase	2.85
DAP1_3085	10523	CHERP	calcium homeostasis endoplasmic reticulum protein	2.81
DAP3_3881	9734	HDAC9	histone deacetylase 9	2.81
DAP3_1658	51406	NOL7	nucleolar protein 7, 27kDa	2.80
DAP3_0609	8826	IQGAP1	IQ motif containing GTPase activating protein 1	2.79
DAP3_3158	120	ADD3	adducin 3 (gamma)	2.79
DAP3_3933	306	ANXA3	annexin A3	2.78
DAP2_5915	10362	HMG20B	high-mobility group 20B	2.76
DAP1_0174	6558	SLC12A2	solute carrier family 12 (sodium/potassium/chloride transporters), member	2.75
DAP2_3448	1282	COL4A1	collagen, type IV, alpha 1	2.75
DAP4_3126	1359	CPA3	carboxypeptidase A3 (mast cell)	2.74
DAP3_1093	3855	KRT7	keratin 7	2.74
DAP1_1741	5271	SERPINB8	serpin peptidase inhibitor, clade B (ovalbumin), member 8	2.69
DAP3_1042	4791	NFKB2	nuclear factor of kappa light polypeptide gene enhancer in B-cells 2	2.67
DAP3_5816	165	AEBP1	AE binding protein 1	2.67
DAP3_3879	7041	TGFB1I1	transforming growth factor beta 1 induced transcript 1	2.66
DAP1_0509	2013	EMP2	epithelial membrane protein 2	2.63
DAP2_3497	596	BCL2	B-cell CLL/lymphoma 2	2.63
DAP3_2152	5698	PSMB9	proteasome (prosome, macropain) subunit, beta type, 9 (large multifunctional peptidase 2)	2.59
DAP3_6062	10097	ACTR2	ARP2 actin-related protein 2 homolog (yeast)	2.59
DAP1_6137	780	DDR1	discoidin domain receptor family, member 1	2.58
DAP2_3913	6541	SLC7A1	solute carrier family 7 (cationic amino acid transporter, y+ system),	2.56
DAP4_2003	5420	PODXL	podocalyxin-like	2.56
DAP1_5750	1307	COL16A1	collagen, type XVI, alpha 1	2.55
DAP1_3284	10437	IFI30	interferon, gamma-inducible protein 30	2.55
DAP3_1596	9852	EPM2AIP1	EPM2A (laforin) interacting protein 1	2.55
DAP3_1678	301	ANXA1	annexin A1	2.53
DAP3_4123	6366	CCL21	chemokine (C-C motif) ligand 21	2.47
DAP3_1610	22856	CHSY1	carbohydrate (chondroitin) synthase 1	2.45
DAP1_4020	162	AP1B1	adaptor-related protein complex 1, beta 1 subunit	2.45
DAP4_2797	7004	TEAD4	TEA domain family member 4	2.39
DAP4_2406	54898	ELOVL2	elongation of very long chain fatty acids (FEN1/Elo2, SUR4/Elo3, yeast)-	2.39
DAP1_0054	6925	TCF4	transcription factor 4	2.38
DAP3_1020	9819	TSC22D2	TSC22 domain family, member 2	2.38
DAP4_2418	1847	DUSP5	dual specificity phosphatase 5	2.36

DAP3_5242	8030	CCDC6	coiled-coil domain containing 6	2.36
DAP3_0973	962	CD48	CD48 molecule	2.35
DAP1_0901	10188	TNK2	tyrosine kinase, non-receptor, 2	2.35
DAP3_1032	1601	DAB2	disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)	2.35
DAP2_3941	4017	LOXL2	lysyl oxidase-like 2	2.34
DAP3_2205	6035	RNASE1	ribonuclease, RNase A family, 1 (pancreatic)	2.34
DAP4_2160	4026	LPP	LIM domain containing preferred translocation partner in lipoma	2.33
DAP3_0038	7852	CXCR4	chemokine (C-X-C motif) receptor 4	2.33
DAP3_1608	6586	SLIT3	slit homolog 3 (Drosophila)	2.31
DAP3_0744	11259	FILIP1L	filamin A interacting protein 1-like	2.25
DAP4_5839	6363	CCL19	chemokine (C-C motif) ligand 19	2.23
DAP3_5744	11214	AKAP13	A kinase (PRKA) anchor protein 13	2.23

Genes correlated with good survival

Probe ID	GeneID	Gene symbol	Description	Cox score
DAP3_4190	223	ALDH9A1	aldehyde dehydrogenase 9 family, member A1	-3.34
DAP4_0296	7276	TTR	transthyretin (prealbumin, amyloidosis type I)	-3.27
DAP1_5588	6018	RLF	rearranged L-myc fusion	-3.23
DAP4_3479	3612	IMPA1	inositol(myo)-1(or 4)-monophosphatase 1	-3.22
DAP3_2208	5207	PFKFB1	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1	-3.22
DAP3_1951	6296	ACSM3	acyl-CoA synthetase medium-chain family member 3	-3.21
DAP4_2813	151	ADRA2B	adrenergic, alpha-2B-, receptor	-3.19
DAP1_3979	5771	PTPN2	protein tyrosine phosphatase, non-receptor type 2	-3.12
DAP3_1558	5691	PSMB3	proteasome (prosome, macropain) subunit, beta type, 3	-3.09
DAP3_2216	5502	PPP1R1A	protein phosphatase 1, regulatory (inhibitor) subunit 1A	-3.07
DAP3_0210	27346	TMEM97	transmembrane protein 97	-3.06
DAP2_4247	5313	PKLR	pyruvate kinase, liver and RBC	-3.01
DAP3_2434	9252	RPS6KA5	ribosomal protein S6 kinase, 90kDa, polypeptide 5	-3.00
DAP1_0453	1528	CYB5A	cytochrome b5 type A (microsomal)	-2.96
DAP4_3541	6447	SCG5	secretogranin V (7B2 protein)	-2.93
DAP1_1650	25828	TXN2	thioredoxin 2	-2.90
DAP2_1608	5340	PLG	plasminogen	-2.88
DAP3_2733	6309	SC5DL	sterol-C5-desaturase (ERG3 delta-5-desaturase homolog, S. cerevisiae)-	-2.87
DAP4_3933	367	AR	androgen receptor (dihydrotestosterone receptor; testicular feminization; spinal and bulbar muscular atrophy; Kennedy disease)	-2.84
DAP3_5880	3479	IGF1	insulin-like growth factor 1 (somatomedin C)	-2.84
DAP1_1983	8802	SUCLG1	succinate-CoA ligase, GDP-forming, alpha subunit	-2.84
DAP3_5885	23498	HAAO	3-hydroxyanthranilate 3,4-dioxygenase	-2.83
DAP2_6048	735	C9	complement component 9	-2.83
DAP4_1959	9013	TAF1C	TATA box binding protein (TBP)-associated factor, RNA polymerase I, C, 110kDa	-2.82
DAP4_2356	1371	CPOX	coproporphyrinogen oxidase	-2.82
DAP4_5179	7507	XPA	xeroderma pigmentosum, complementation group A	-2.82
DAP4_0915	3026	HABP2	hyaluronan binding protein 2	-2.81
DAP3_3625	2690	GHR	growth hormone receptor	-2.77
DAP4_1564	5105	PCK1	phosphoenolpyruvate carboxykinase 1 (soluble)	-2.76
DAP2_1588	6718	AKR1D1	aldo-keto reductase family 1, member D1 (delta 4-3-ketosteroid-5-beta-	-2.76
DAP3_1407	128	ADH5	alcohol dehydrogenase 5 (class III), chi polypeptide	-2.75
DAP3_5846	16	AARS	alanyl-tRNA synthetase	-2.70
DAP4_1895	732	C8B	complement component 8, beta polypeptide	-2.69
DAP1_2114	51237	MGC29506	NA	-2.67
DAP4_3262	10159	ATP6AP2	ATPase, H+ transporting, lysosomal accessory protein 2	-2.67
DAP4_2906	9732	DOCK4	dedicator of cytokinesis 4	-2.66
DAP4_4262	5627	PROS1	protein S (alpha)	-2.66
DAP4_5591	7709	ZBTB17	zinc finger and BTB domain containing 17	-2.65
DAP1_2989	1603	DAD1	defender against cell death 1	-2.65
DAP4_0781	1678	TIMM8A	translocase of inner mitochondrial membrane 8 homolog A (yeast)	-2.65
DAP3_5291	3155	HMGCL	3-hydroxymethyl-3-methylglutaryl-Coenzyme A lyase (hydroxymethylglutaricaciduria)	-2.65
DAP3_4919	725	C4BPB	complement component 4 binding protein, beta	-2.62
DAP4_5846	7189	TRAF6	TNF receptor-associated factor 6	-2.62
DAP1_0147	1967	EIF2B1	eukaryotic translation initiation factor 2B, subunit 1 alpha, 26kDa	-2.61
DAP1_0559	3990	LIPC	lipase, hepatic	-2.60
DAP4_5383	10026	PIGK	phosphatidylinositol glycan anchor biosynthesis, class K	-2.60
DAP4_5653	80344	WDR23	WD repeat domain 23	-2.59
DAP4_0010	5982	RFC2	replication factor C (activator 1) 2, 40kDa	-2.58
DAP4_5452	2915	GRM5	glutamate receptor, metabotropic 5	-2.56
DAP3_1646	6391	SDHC	succinate dehydrogenase complex, subunit C, integral membrane protein,	-2.55

DAP3_2354	2073	ERCC5	excision repair cross-complementing rodent repair deficiency, complementation group 5 (xeroderma pigmentosum, complementation group G (Cockayne syndrome))	-2.54
DAP1_2179	2158	F9	coagulation factor IX (plasma thromboplastic component, Christmas disease, hemophilia B)	-2.54
DAP2_2062	157567	ANKRD46	ankyrin repeat domain 46	-2.54
DAP3_2994	417	ART1	ADP-ribosyltransferase 1	-2.54
DAP3_1761	1486	CTBS	chitobiase, di-N-acetyl-	-2.54
DAP3_3022	2542	SLC37A4	solute carrier family 37 (glucose-6-phosphate transporter), member 4	-2.53
DAP4_3697	211	ALAS1	aminolevulinate, delta-, synthase 1	-2.53
DAP4_5013	27072	VPS41	vacuolar protein sorting 41 homolog (S. cerevisiae)	-2.51
DAP3_1312	2642	GCGR	glucagon receptor	-2.51
DAP1_5069	10694	CCT8	chaperonin containing TCP1, subunit 8 (theta)	-2.51
DAP1_0656	25874	BRP44	brain protein 44	-2.50
DAP1_5381	2868	GRK4	G protein-coupled receptor kinase 4	-2.50
DAP4_1861	3336	HSPE1	heat shock 10kDa protein 1 (chaperonin 10)	-2.50
DAP2_5268	79731	NARS2	asparaginyl-tRNA synthetase 2, mitochondrial (putative)	-2.49
DAP1_5672	667	DST	dystonin	-2.49
DAP1_5518	27032	ATP2C1	ATPase, Ca++ transporting, type 2C, member 1	-2.48
DAP4_3497	10327	AKR1A1	aldo-keto reductase family 1, member A1 (aldehyde reductase)	-2.48
DAP1_1085	2010	EMD	emerin (Emery-Dreifuss muscular dystrophy)	-2.47
DAP4_5050	799	CALCR	calcitonin receptor	-2.45
DAP3_4223	22839	DLGAP4	discs, large (Drosophila) homolog-associated protein 4	-2.45
DAP4_3111	6240	RRM1	ribonucleotide reductase M1 polypeptide	-2.44
DAP4_3810	29937	NENF	neuron derived neurotrophic factor	-2.44
DAP1_3440	29887	SNX10	sorting nexin 10	-2.44
DAP3_5257	5372	PMM1	phosphomannomutase 1	-2.44
DAP1_5842	6999	TDO2	tryptophan 2,3-dioxygenase	-2.43
DAP4_3363	2944	GSTM1	glutathione S-transferase M1	-2.43
DAP1_5123	6721	SREBF2	sterol regulatory element binding transcription factor 2	-2.42
DAP4_0140	26469	PTPN18	protein tyrosine phosphatase, non-receptor type 18 (brain-derived)	-2.42
DAP3_1623	27163	ASAHL	N-acylsphingosine amidohydrolase (acid ceramidase)-like	-2.41
DAP2_4928	5336	PLCG2	phospholipase C, gamma 2 (phosphatidylinositol-specific)	-2.41
DAP3_5959	3760	KCNJ3	potassium inwardly-rectifying channel, subfamily J, member 3	-2.40
DAP3_1753	5833	PCYT2	phosphate cytidylyltransferase 2, ethanolamine	-2.40
DAP4_4304	2705	GJB1	gap junction protein, beta 1, 32kDa	-2.39
DAP3_5067	7108	TM7SF2	transmembrane 7 superfamily member 2	-2.39
DAP4_5379	8991	SELENBP1	selenium binding protein 1	-2.38
DAP4_3066	316	AOX1	aldehyde oxidase 1	-2.37
DAP3_2882	10444	ZER1	zer-1 homolog (C. elegans)	-2.37
DAP4_6012	130	ADH6	alcohol dehydrogenase 6 (class V)	-2.36
DAP3_5076	2956	MSH6	mutS homolog 6 (E. coli)	-2.36
DAP2_3569	8671	SLC4A4	solute carrier family 4, sodium bicarbonate cotransporter, member 4	-2.34
DAP3_3988	9097	USP14	ubiquitin specific peptidase 14 (tRNA-guanine transglycosylase)	-2.34
DAP3_6123	727	C5	complement component 5	-2.32
DAP4_0949	5893	RAD52	RAD52 homolog (S. cerevisiae)	-2.32
DAP4_0979	116496	FAM129A	family with sequence similarity 129, member A	-2.31
DAP4_2296	10458	BAIAP2	BAI1-associated protein 2	-2.31
DAP1_1550	6744	SSFA2	sperm specific antigen 2	-2.30
DAP2_6140	5446	PON3	paraoxonase 3	-2.30
DAP3_2198	2646	GCKR	glucokinase (hexokinase 4) regulator	-2.30
DAP3_3783	1385	CREB1	cAMP responsive element binding protein 1	-2.30
DAP3_3049	23316	CUTL2	cut-like 2 (Drosophila)	-2.29
DAP1_5546	6427	SFRS2	splicing factor, arginine/serine-rich 2	-2.28
DAP4_0984	3156	HMGCR	3-hydroxy-3-methylglutaryl-Coenzyme A reductase	-2.28
DAP3_5468	2677	GGCX	gamma-glutamyl carboxylase	-2.27
DAP2_5898	1555	CYP2B6	cytochrome P450, family 2, subfamily B, polypeptide 6	-2.26
DAP4_3279	7739	ZNF185	zinc finger protein 185 (LIM domain)	-2.26
DAP3_1562	378	ARF4	ADP-ribosylation factor 4	-2.23
DAP4_3503	10965	ACOT2	acyl-CoA thioesterase 2	-2.22
DAP3_0889	513	ATP5D	ATP synthase, H+ transporting, mitochondrial F1 complex, delta subunit	-2.22
DAP2_4148	1369	CPN1	carboxypeptidase N, polypeptide 1	-2.20
DAP2_1935	5331	PLCB3	phospholipase C, beta 3 (phosphatidylinositol-specific)	-2.20
DAP3_2137	3642	INSM1	insulinoma-associated 1	-2.18
DAP4_3027	5442	POLRMT	polymerase (RNA) mitochondrial (DNA directed)	-2.14
DAP3_5700	11145	HRASLS3	HRAS-like suppressor 3	-2.13

Supplementary Table 3
Functional annotation of Survival signature by Gene Set Enrichment Analysis (Training set)
(For details of each gene set, click the name for the link to MSigDB gene set annotation page)

(a) Gene sets correlated with poor survival

Experimental perturbation gene set	# genes	NES	FDR
IFNA_HCMV_6HRS_UP	56	2.30	0.000
CROONQUIST_IL6_STROMA_UP	34	2.19	0.003
SANA_IFNG_ENDOTHELIAL_UP	30	1.98	0.034
IFN_ALPHA_UP	30	1.96	0.029
SANA_TNFA_ENDOTHELIAL_UP	46	1.96	0.024
RADAEVA_IFNA_UP	29	1.96	0.020
ADIP_HUMAN_DN	18	1.95	0.018
IFNA_UV-CMV_COMMON_HCMV_6HRS_UP	33	1.94	0.019
O6BG_RESIST_MEDULLOBLASTOMA_DN	24	1.90	0.025
HADDAD_HSC_CD10_UP	165	1.86	0.034
TGFBETA_ALL_UP	50	1.85	0.035
ZUCCHI_EPITHELIAL_DN	34	1.83	0.038
BRG1_ALAB_DN	18	1.83	0.036
CROONQUIST_RAS_STROMA_DN	18	1.81	0.041
HINATA_NFKB_UP	91	1.71	0.087

Litrature-based pathway gene set	# genes	NES	FDR
INFLAMMATORY_RESPONSE_PATHWAY	25	1.90	0.023

(b) Gene sets correlated with good survival

Experimental perturbation gene set	# genes	NES	FDR
FETAL_LIVER_VS_ADULT_LIVER_GNF2	44	-2.20	0.002

Litrature-based pathway gene set	# genes	NES	FDR
ANDROGEN_AND_ESTROGEN_METABOLISM	21	-2.17	0.001
FATTY_ACID_METABOLISM	57	-2.15	0.001
TRYPTOPHAN_METABOLISM	44	-2.08	0.002
BILE_ACID_BIOSYNTHESIS	17	-2.02	0.004
ELECTRON_TRANSPORT_CHAIN	48	-2.01	0.003
VALINE_LEUCINE_AND_ISOLEUCINE_DEGRADATION	23	-1.99	0.003
INOSITOL_PHOSPHATE_METABOLISM	19	-1.90	0.007
BUTANOATE_METABOLISM	20	-1.89	0.006
BETA_ALANINE_METABOLISM	21	-1.81	0.014
PYRUVATE_METABOLISM	26	-1.75	0.023
GLYCINE_SERINE_AND_THREONINE_METABOLISM	23	-1.74	0.023
GAMMA_HEXACHLOROCYCLOHEXANE_DEGRADATION	25	-1.71	0.028
GLYCEROLIPID_METABOLISM	27	-1.70	0.028

NES: normalized enrichment score, FDR: false discovery rate

Supplementary Table 4

Gene expression-based survival prediciton and histological inflammation of the liver (Training set)

	Inflammation			
Prediction	None	Mild	Moderate	Severe
Poor survival	2	11	9	5
Good survival	4	17	22	11

Fisher's exact test, p=0.89

Scored according to Batts K, Ludwig J. Am J Surg Pathol 19:1409,1995

Supplementary Table 5
Univariate Cox regression analysis of clinical risk factors (Validation set)

Early recurrence

Variable	Hazard ratio	95% CI		p-value
		low	high	
Multinodularity	1.95	1.13	3.37	0.02
Vascular invasion	1.72	1.11	2.66	0.02
AFP > 100 ng/mL	1.94	1.21	3.12	0.006

Late recurrence

Variable	Hazard ratio	95% CI		p-value
		low	high	
Multinodularity	2.10	0.73	6.06	0.17
Vascular invasion	0.85	0.38	1.92	0.70
AFP > 100 ng/mL	0.45	0.17	1.19	0.11

Survival

Variable	Hazard ratio	95% CI		p-value
		low	high	
Multinodularity	1.66	0.77	3.59	0.19
Vascular invasion	2.05	1.12	3.76	0.02
AFP > 100 ng/mL	2.10	1.08	4.07	0.03

AFP: alpha-fetoprotein

Supplementary Table 6
Multivariate Cox regression: subgroup analysis (Validation set)

Late recurrence (longer follow-up patients, n=167)

Variable	Hazard ratio	95% CI		p-value
		low	high	
Late recurrence signature	2.94	1.39	6.20	0.005

Late recurrence (BCLC ≤ A, n=207)

Variable	Hazard ratio	95% CI		p-value
		low	high	
Late recurrence signature	2.97	1.37	6.45	0.006

Survival (BCLC ≤ A, n=207)

Variable	Hazard ratio	95% CI		p-value
		low	high	
Survival signature	1.93	0.87	4.28	0.10
AFP > 100 ng/mL	2.30	1.04	5.05	0.04
Vascular invasion	1.80	0.84	3.88	0.13

Survival (longer follow-up patients, BCLC ≤ A, n=154)

Variable	Hazard ratio	95% CI		p-value
		low	high	
Survival signature	2.04	0.91	4.59	0.08
AFP > 100 ng/mL	2.13	0.95	4.76	0.07
Vascular invasion	2.01	0.92	4.36	0.08

AFP: alpha-fetoprotein

Supplementary Table 7A
Clonality analysis of paired primary and recurrent HCC

Case ID	"Heterozygous in recurrence" / "Homozygous in primary"	"Homozygous in recurrence" / "Heterozygous in primary"	Primary tumor subclass*	Recurrent tumor subclass*
hcc_018	26% (562/2167)	25% (268/1086)	S2	S1
hcc_044	6% (178/2887)	13% (156/1200)	S1	S3
hcc_082	1% (31/3244)	10% (123/1205)	S3	S2
hcc_075	9% (218/2548)	15% (171/1168)	S2	S1
hcc_101	8% (168/2063)	32% (310/967)	S2	S2
hcc_104	-	-	S3	S1

*Molecular subclasses of HCC defined by a meta-analysis of published frozen sample-based microarray datasets (Hoshida et al. Manuscript in preparation).
"Heterozygous in recurrence" / "Homozygous in primary" in adjacent non-tumor liver tissues of hcc_082, hcc_075, and Hela cells were 0.1% (4/3759), 0.6% (20/3275), and 0.3% (10/3701), respectively.
"Homozygous in recurrence" / "Heterozygous in primary" in adjacent non-tumor liver tissues of hcc_082, hcc_075, and Hela cells were 0.6% (11/1869), 2.3% (38/1676), and 0.7% (8/1216), respectively.

Supplementary Table 7B
Clonality analysis of paired primary and recurrent/metastatic non-HCC tumors

Cancer type	"Heterozygous in recurrence" / "Homozygous in primary"	"Homozygous in recurrence" / "Heterozygous in primary"
Endometrial 1	0.1% (19/40897)	0.8% (124/15033)
Endometrial 2	1.9% (1621/83369)	0.6% (160/24907)
Endometrial 3	0.2% (198/82423)	0.9% (229/26372)
Ovarian 1	1.9% (1686/90069)	1.7% (300/17747)
Ovarian 2	1.3% (1182/91547)	3.3% (549/16621)
Ovarian 3	0.3% (121/44828)	2.2% (218/10075)
Ovarian 4	0.2% (103/43429)	4.5% (441/9781)
Renal 1	1.6% (682/42699)	0.1% (13/13701)
Renal 2	1.9% (796/42753)	0.1% (9/12720)
Renal 3	0.1% (55/42518)	1.2% (157/13154)
DLBCL 1	1.1% (6503/605724)	2.8% (6001/216597)
DLBCL 2	0.9% (5361/598243)	2.2% (5089/228143)
DLBCL 3	0.4% (2216/625092)	0.6% (1402/239662)
DLBCL 4	1.0% (6377/618375)	2.0% (4525/222075)
DLBCL 5	1.8% (10689/582552)	2.5% (5697/232207)
DLBCL 6	0.6% (3773/600753)	6.1% (13368/219845)

DLBCL: diffuse large B-cell lymphoma
Endometrial, ovarian, and renal cancers were profiles on Afymetrix 500k SNP array.
DLBCL samples were profiled on Affymetrix SNP 6.0 array.

Supplementary Table 8
Datasets used to select Transcriptionally Informative Genes

Tissue type	Disease	# samples	Reference
Brain	Glioblastoma	50	Cancer Res 63;1602,2003
	Medulloblastoma	60	Nature 415;436,2002
	Medulloblastoma	23	Nature Genet 29;143,2001
Breast	Breast cancer	73	Unpublished
	Breast cancer	49	PNAS 98;11462,2001, Lancet 361;1590,2003
	Breast cancer	40	Unpublished
Lung	Lung cancer	62	PNAS 98;13790,2001
	Lung cancer	86	Nature Med 8;816,2002
Stomach	Gastric cancer	30	Cancer Res 62;233,2003
Liver	Hepatocellular carcinoma	49	Cancer Res 64;7263,2004
	Hepatocellular carcinoma	60	Lancet 361;923,2003
Ovary	Ovarian cancer	113	Unpublished
Prostate	Prostate cancer	102	Cacer Cell 1;203,2002
	Prostate cancer	120	Unpublished
	Prostate cancer	80	Unpublished
Hematopoetic	Diffuse large B-cell lymphoma	176	Blood 105;1851,2005
	Diffuse large B-cell lymphoma	210	Blood 102;3871,2003
	Acute myeloid/lymphoblastic leuken	52	Unpublished
	Mixed-lineage leukemia	72	Nature Genet 30;41,2002
Skin	Melanoma	115	Unpublished
Astrocyte	Astrocytoma	13	Cancer Res 63;1865,2003
Cancer & normal tissues*	Cancer & normal tissues	280	PNAS 98;15149,2001
Cancer tissues*	Primary & metastatic cancers	76	Nature Genet 33;49,2003
Normal tissues*	Normal tissues	158	PNAS 101;6062,2004
		2149	

*: Panel of multiple tissue types

Supplementary Table 9
Concordance in gene expression change (DHL4 vs. Hela cell lines) between intact and FFPE-RNA on DASL assay

	All genes	Fold change in fresh RNA		
		> 2-fold	> 5-fold	> 10-fold
# genes DHL4 > Hela in fresh RNA	3156	811	185	93
# genes with concordant change in FFPE RNA	2282 (72%)	687 (85%)	180 (97%)	91 (98%)
# genes with discordant change in FFPE RNA	874 (28%)	124 (15%)	5 (3%)	2 (2%)
# genes DHL4 < Hela in fresh RNA	2988	1056	339	138
# genes with concordant change in FFPE RNA	2135 (71%)	905 (86%)	321 (95%)	137 (99%)
# genes with discordant change in FFPE RNA	853 (29%)	151 (14%)	18 (5%)	1 (1%)

Supplementary Table 10
Leave-one-out cross-validation error rates for outcome prediction using HCC tissue data (Training set).

Prediction algorithm	Outcome	
	Survival	Recurrence
CART	40%	21%
k-NN, 1neighbor	41%	18%
k-NN, 3 neighbors	43%	18%
k-NN, 5 neighbors	31%	19%
k-NN, 7 neighbors	36%	19%
WV, 10 markers	38%	41%
WV, 50 markers	48%	45%
WV, 100 markers	49%	30%
SVM	43%	23%

CART: classification and regression trees,
k-NN: k-nearest neighbor,
WV: weighted voting,
SVM: support vector machine

Supplementary Figure legends

Supplementary Figure 1

Effect of missing gene expression signals by reducing the number of probes for each gene in the DASL assay. A: Missing signals by reducing the number of probes assigned for each gene. Left panel shows expression levels of 502 cancer-related genes (Cancer Panel, Illumina) computed as average of 3 independent probes for each gene. Right panel shows signals falling below the level of negative control probes (black bars) by randomly picking a single probe from the 3 probes representing each gene. B: Hierarchical clustering using 5 datasets generated by randomly picking 1 probe from the 3 probes. C: Comparison of rank of top HCC marker genes (top and bottom 20 genes) between 1-probe and 3-probe datasets.

Supplementary Figure 2

A: Leave-one-out cross validation-based survival prediction using FFPE HCC tissues. B: Previously reported survival-predictive signature (Lee, et al. Hepatology 40:667,2004) recapitulated in the dataset (left panel) without association with survival (right panel).

Supplementary Figure 3

A: Leave-one-out cross validation-based survival prediction using publicly available gene expression dataset of fresh frozen HCC tissues (n=67, NCBI Gene Expression Omnibus dataset accession # GSE9843). B: Previously reported survival-predictive signature (Lee, et al. Hepatology 40:667,2004) recapitulated in the dataset (left panel) without association with survival (right panel).

Supplementary Figure 4

Smoothed tumor recurrence hazard over time after surgery for training (A) and validation (B) sets. There is no peak of early recurrence in training set.

Supplementary Figure 5

Survival curves according to the grade of hepatitis activity (based on Batts and Ludwig. Am J Surg Pathol 19:1409,1995) in the training set.

Supplementary Figure 6

Overall recurrence curves in the validation set according to the prediction made by the late recurrence-predictive signature (132 genes, A) and the overall recurrence-predictive signature (174 genes, B). C: Correlation between survival- and late recurrence-predictive signatures: genes on microarray were rank-ordered according to their correlation with survival time, and subset of late recurrence signature genes associated with higher (upper panel) or lower (lower panel) risk of late recurrence was separately evaluated for its overrepresentation on poor survival or good survival side in the rank-ordered gene list, respectively, using Gene Set Enrichment Analysis ($p<0.001$, see **Supplementary Appendix**). Early recurrences (< 2 years following resection) are censored in the analysis of late recurrence. Red and blue lines indicate prediction of higher and lower risk of late/overall recurrence, respectively.

Supplementary Figure 7

Assessment of clonality between primary and recurrent tumors. A: The panel shows how many homozygous loci in the primary tumors appear to be heterozygous in paired recurrent tumors. B: The panel shows how many heterozygous loci in the primary tumors appear to be homozygous in paired recurrent tumors. DLBCL: diffuse large B-cell lymphoma.

Supplementary Figure 8

"Present" gene expression signals in genome-wide microarray datasets profiling panels of multiple human tissue types. A: Panel of cancer tissues (PNAS 2001;98:15149, http://www.broad.mit.edu/cancer/). B: Panel of normal tissues (PNAS 2004;101;6062, http://www.gnf.org/). Red color indicates "present" (i.e., expressed) genes.

Supplementary Figure 9

Selection process for 6,000 transcriptionally informative genes in the DASL assay. A: In each of previously generated 24 microarray datasets, coefficient of variation (CV) was calculated for each gene and summarized on to the list of NCBI RefSeq ID. B: The top 6,000 genes cover 70-90% of genes in microarray-based signatures (375 gene

sets) and literature-based molecular pathways (450 gene sets) collected in Molecular Signature Database (MSigDB), C: Age of FFPE blocks and %P-call in 10 prostate cancer samples. Red arrow head indicates samples fixed 24 years before RNA extraction; blue arrow head indicates a sample fixed 7 years before RNA extraction.

Supplementary Figure 10

Quality assessment of DASL profile based on the proportion of "present" (i.e., expressed) genes (%P-call) in the training set. Correlation coefficient of each array to the "median" array was plotted against %P-call for tumor (left) and adjacent liver (right) profiles from the training set. For each tissue type, quality threshold was defined as a %P-call where the correlation starts to drop. Green lines indicate %P-call threshold of 65% and 70% for tumor and liver profiles, respectively. The same quality threshold was applied to the profiles from validation set.

Supplementary Figure 11

Comparison of gene expression fold change between intact and FFPE-RNA.

Supplementary Figure 12

Prediction of prostate cancer using the DASL profile of marker genes defined by a meta-analysis of published 7 frozen sample-based microarray datasets.

Supplementary Figure 13

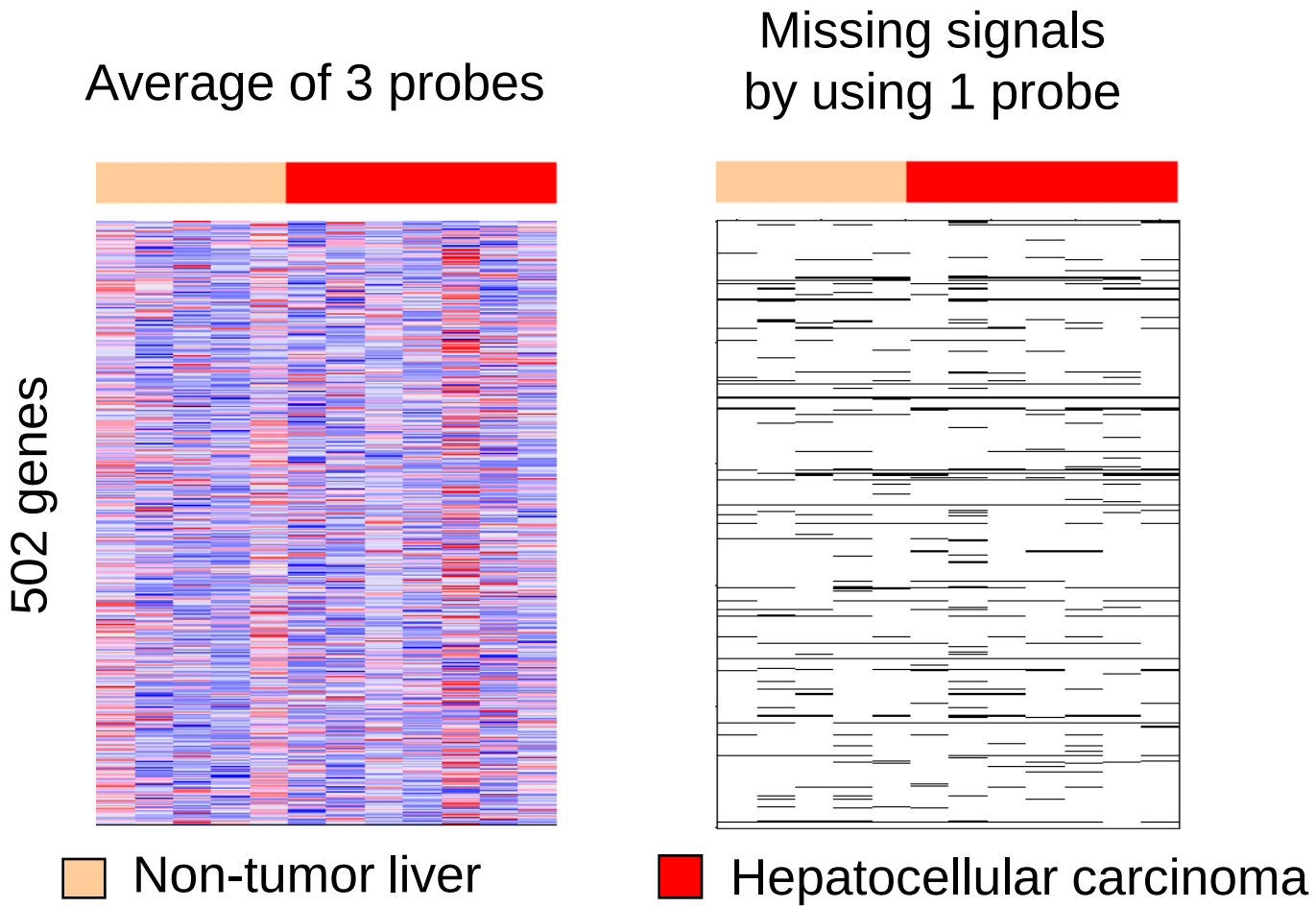
The survival signature in a publicly available independent dataset of fresh frozen non-tumor liver tissues (n=10).

Supplementary Figure 14

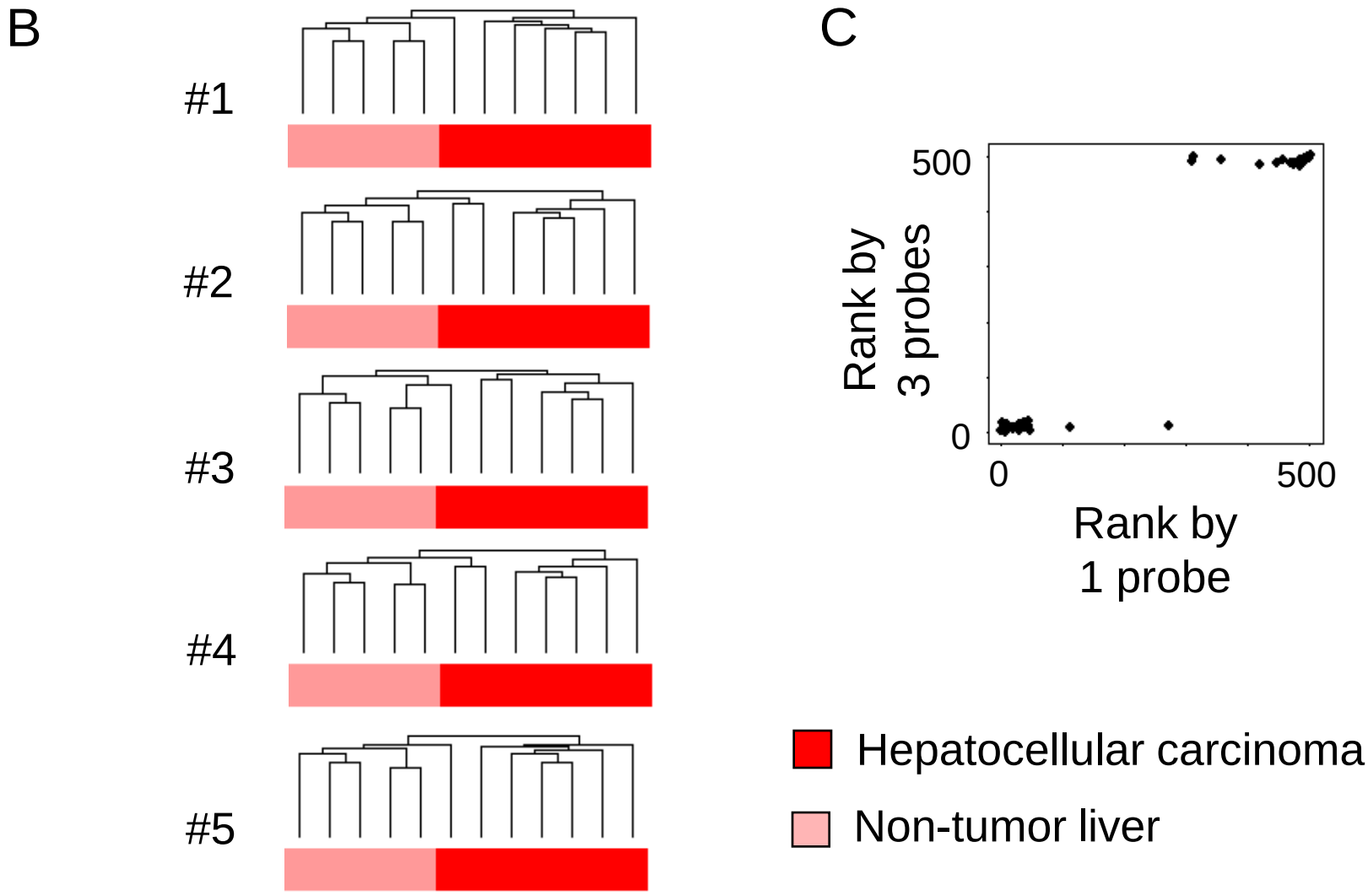
Survival curves for three geographic sites in the validation set: US (n=88, median follow-up 2.4 years), Spain (n=45, median follow-up 3.1 years), and Italy (n=92, median follow-up 1.9 years). A: Overall survival. B: Survival curves according to the survival prediction. Red lines indicate poor survival prediction; blue lines indicate good survival prediction.

Supplementary Figure 1A

A

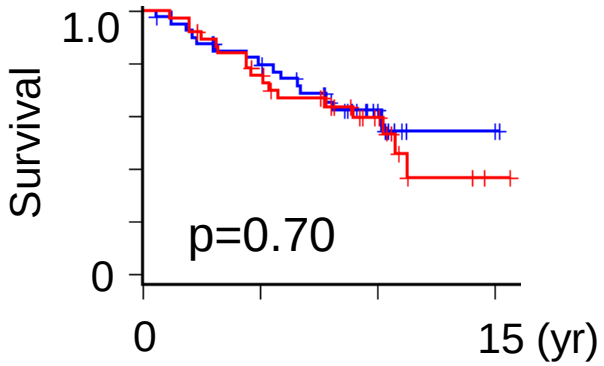


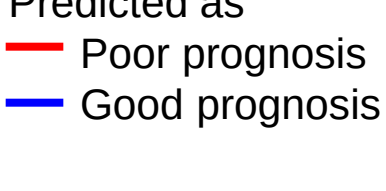
Supplementary Figure 1B,C



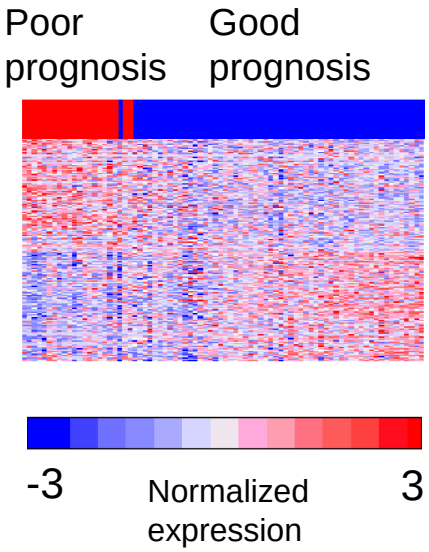
Supplementary Figure 2

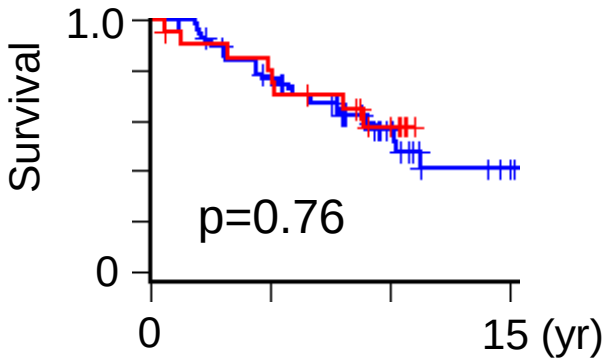
A





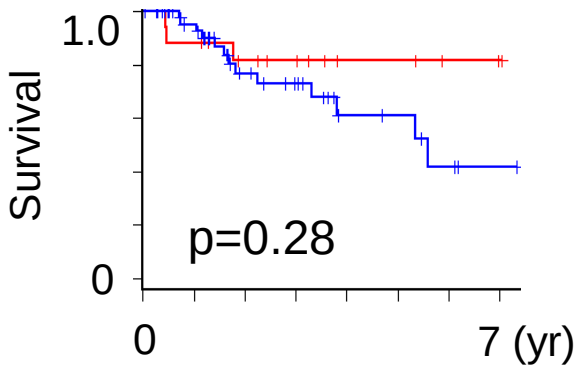
B

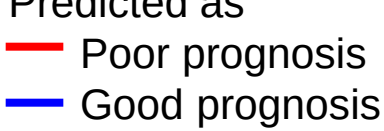




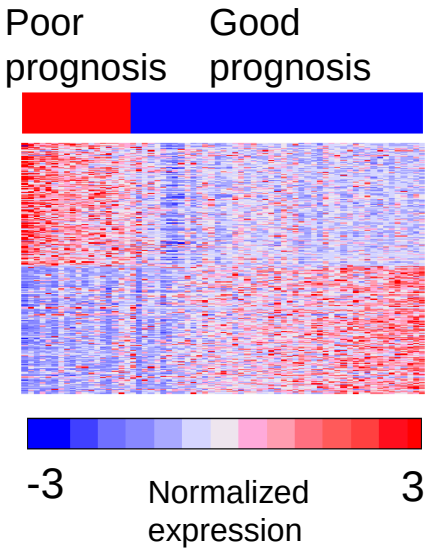
Supplementary Figure 3

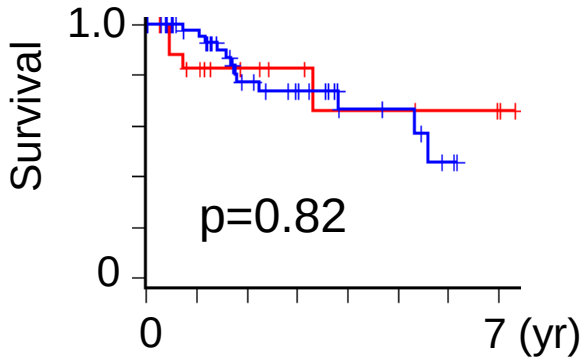
A



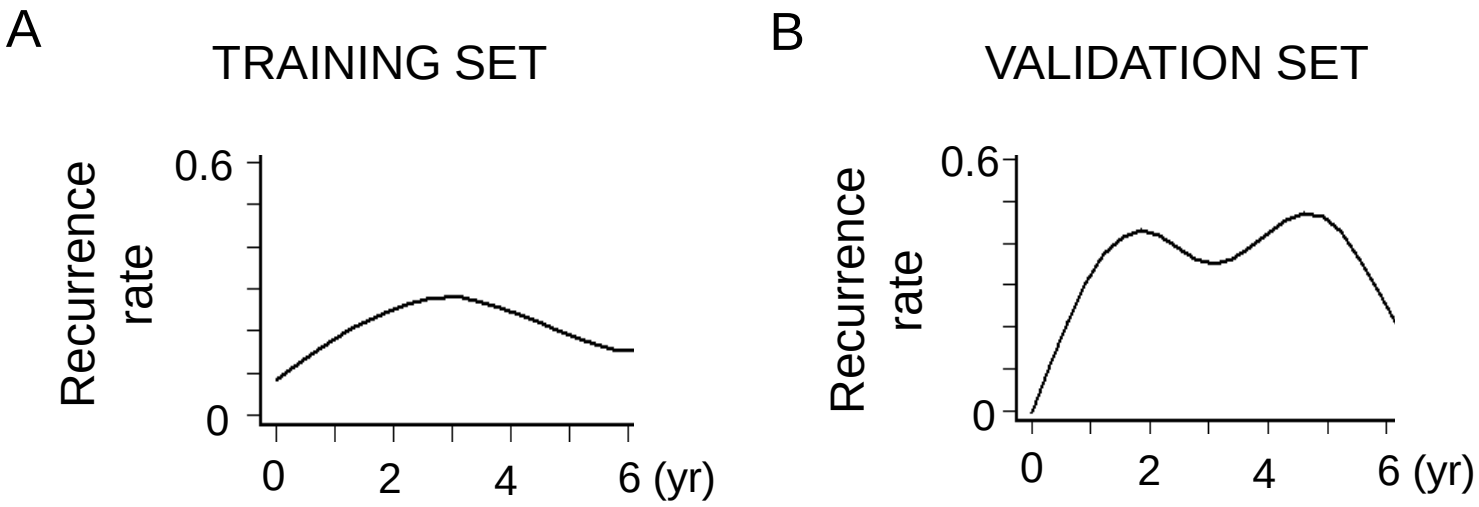


B

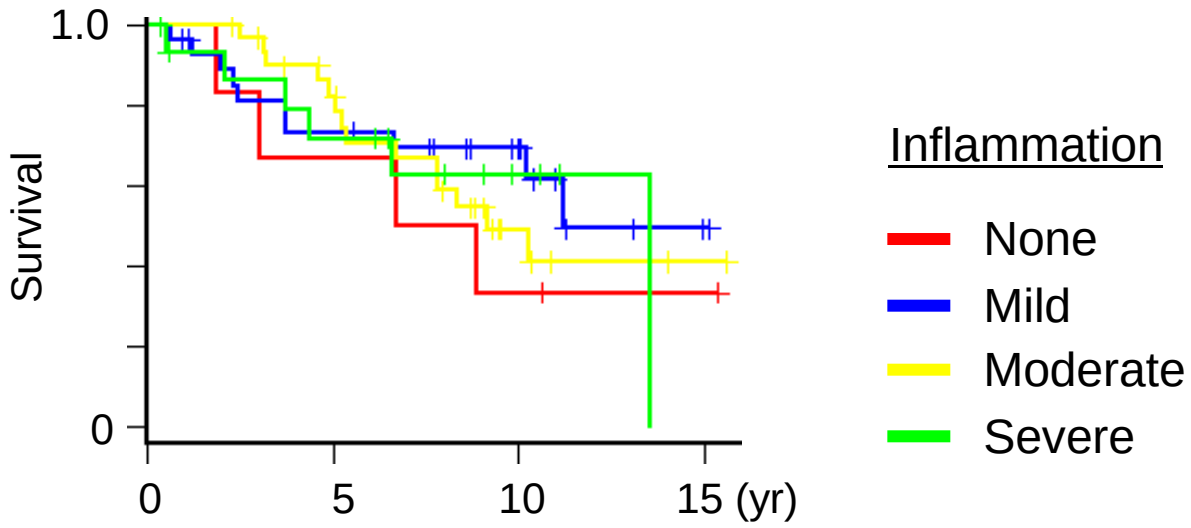




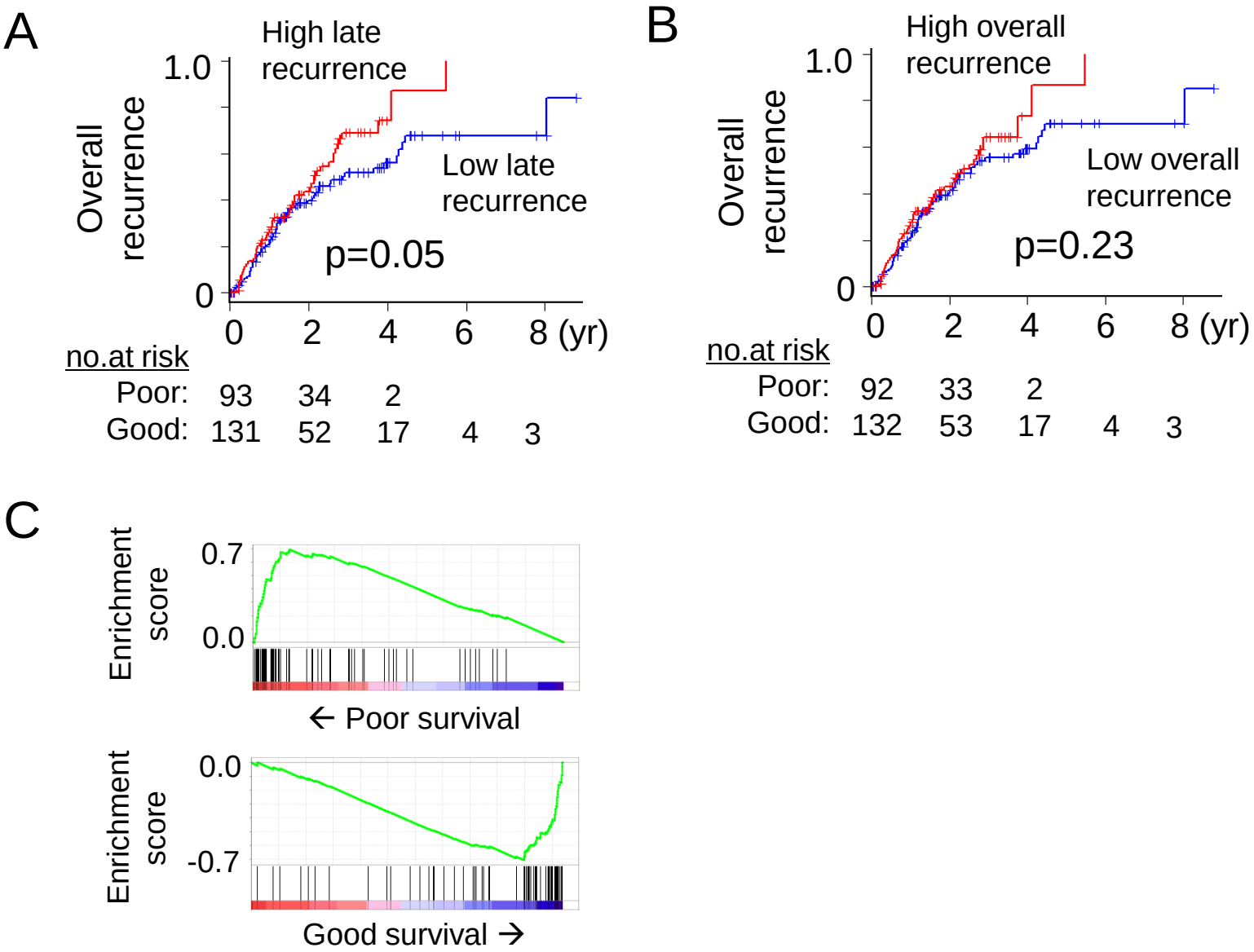
Supplementary Figure 4



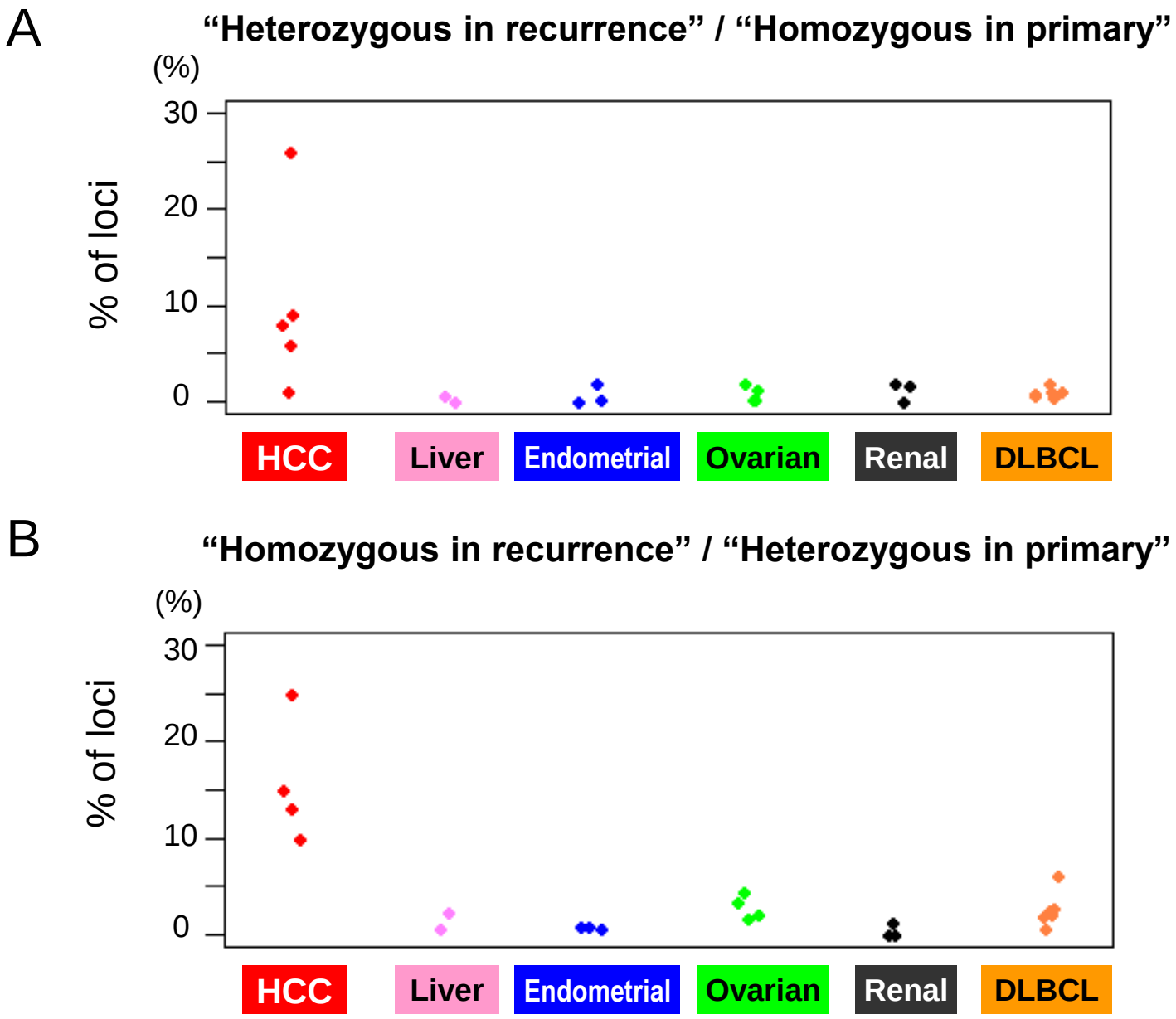
Supplementary Figure 5



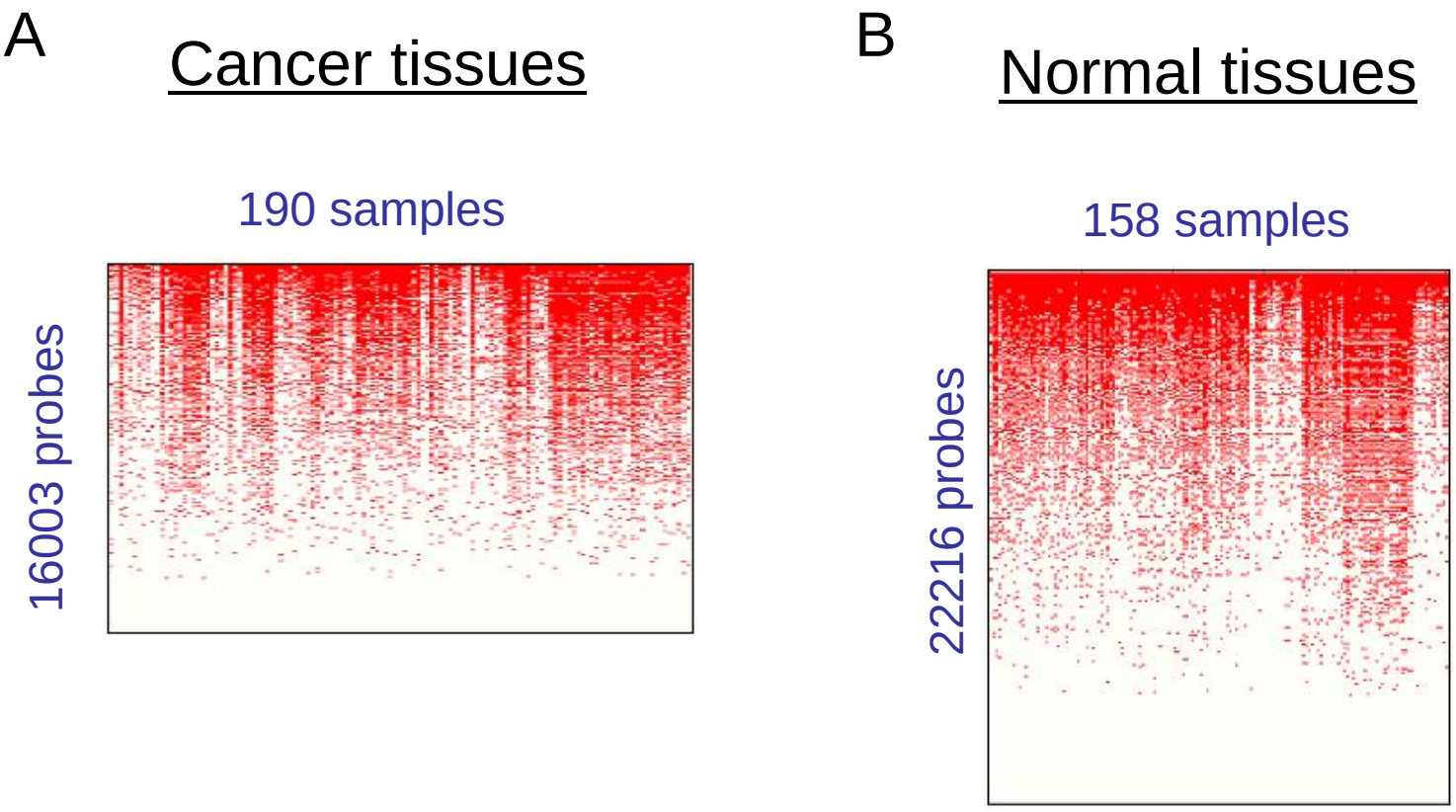
Supplementary Figure 6



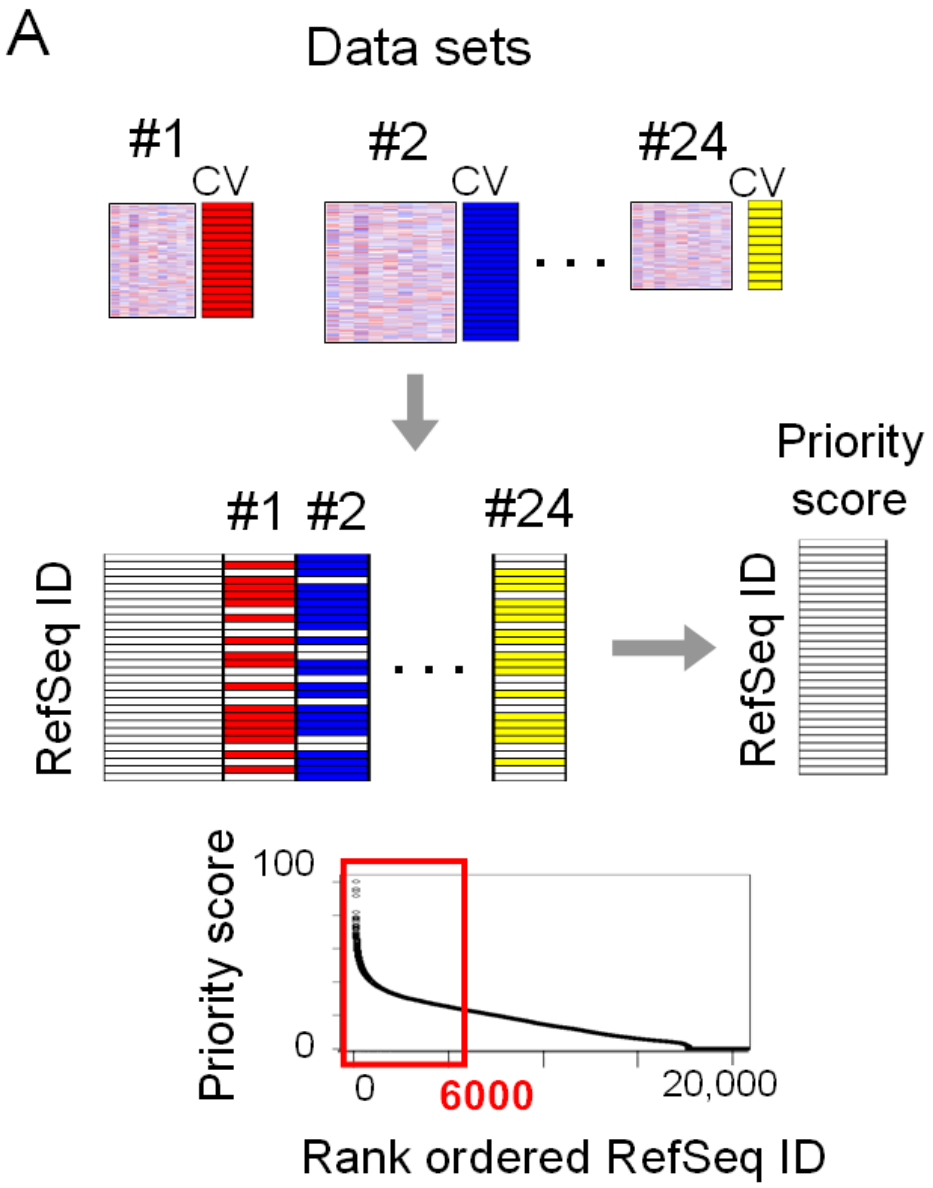
Supplementary Figure 7

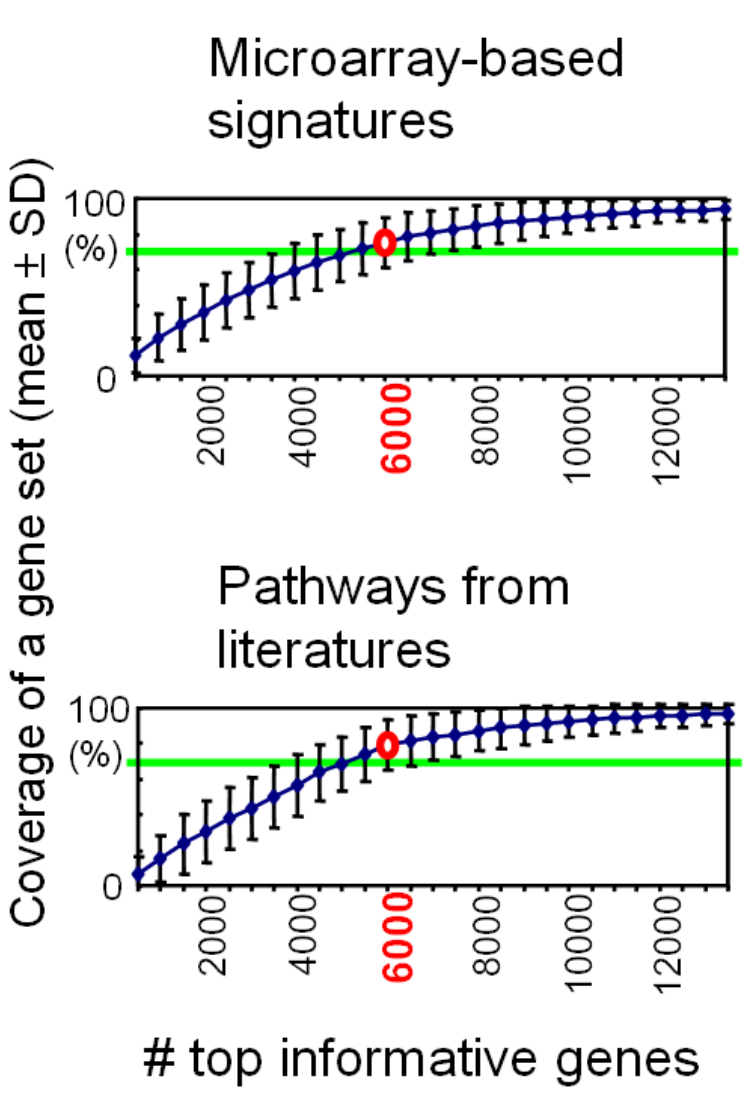


Supplementary Figure 8

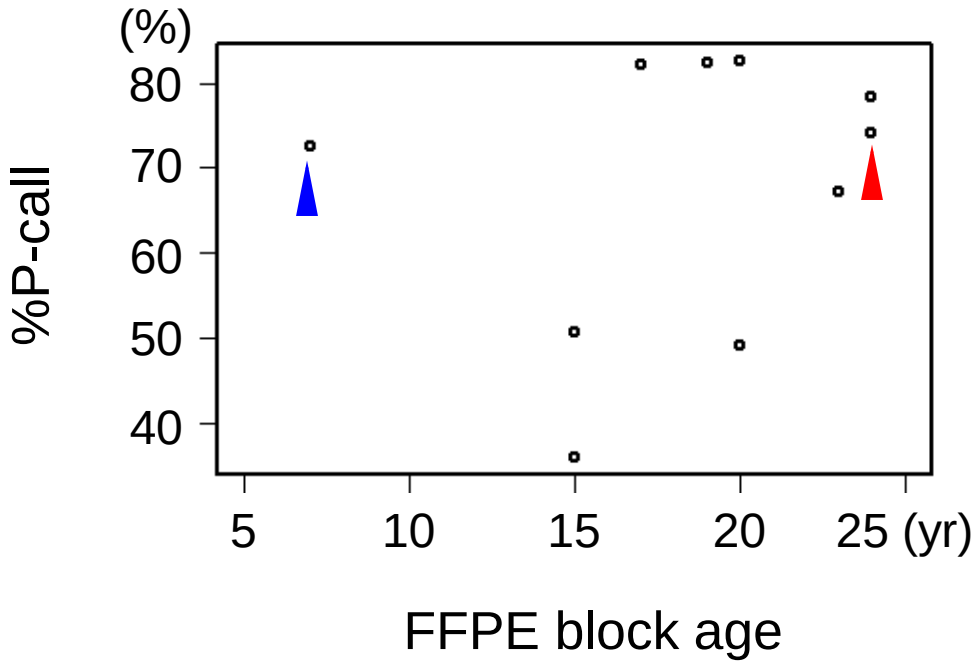


Supplementary Figure 9A,B



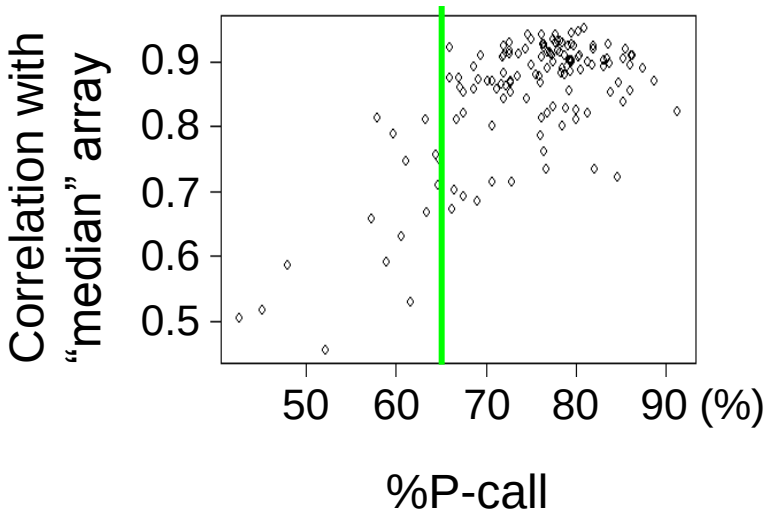


Supplementary Figure 9C

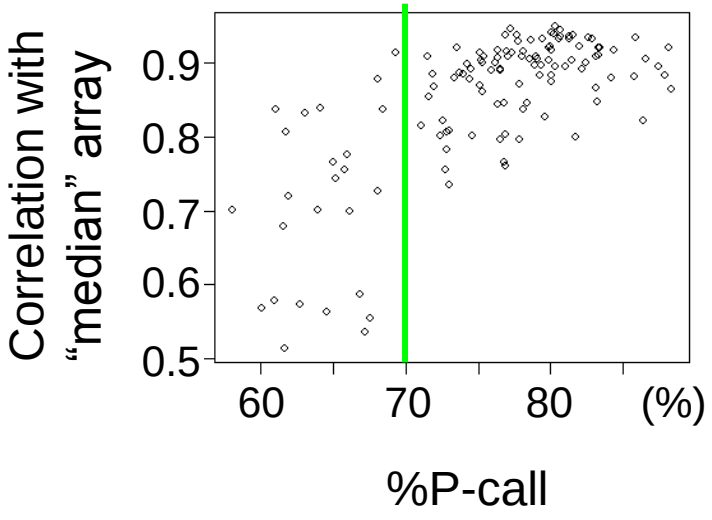


Supplementary Figure 10

A Tumor profiles

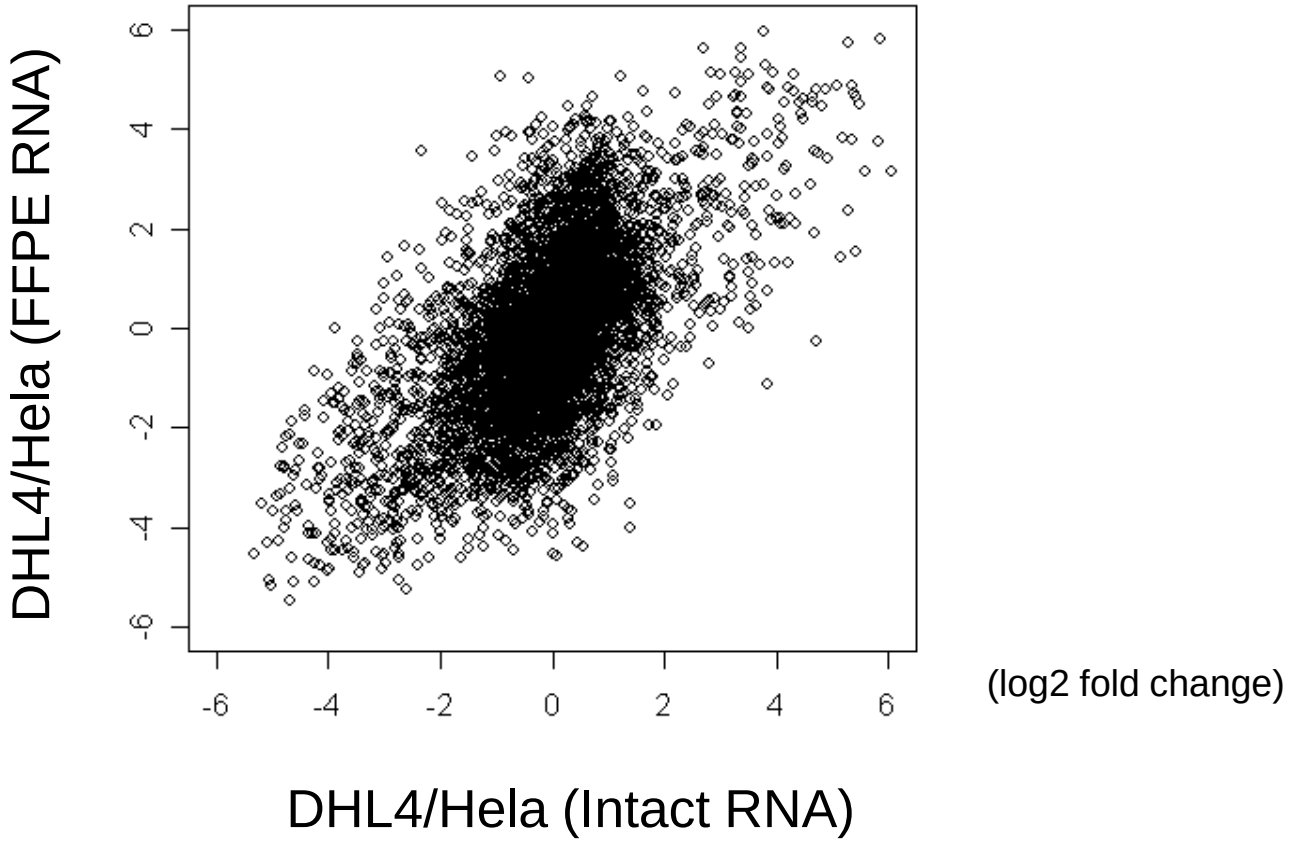


B Adjacent liver profiles



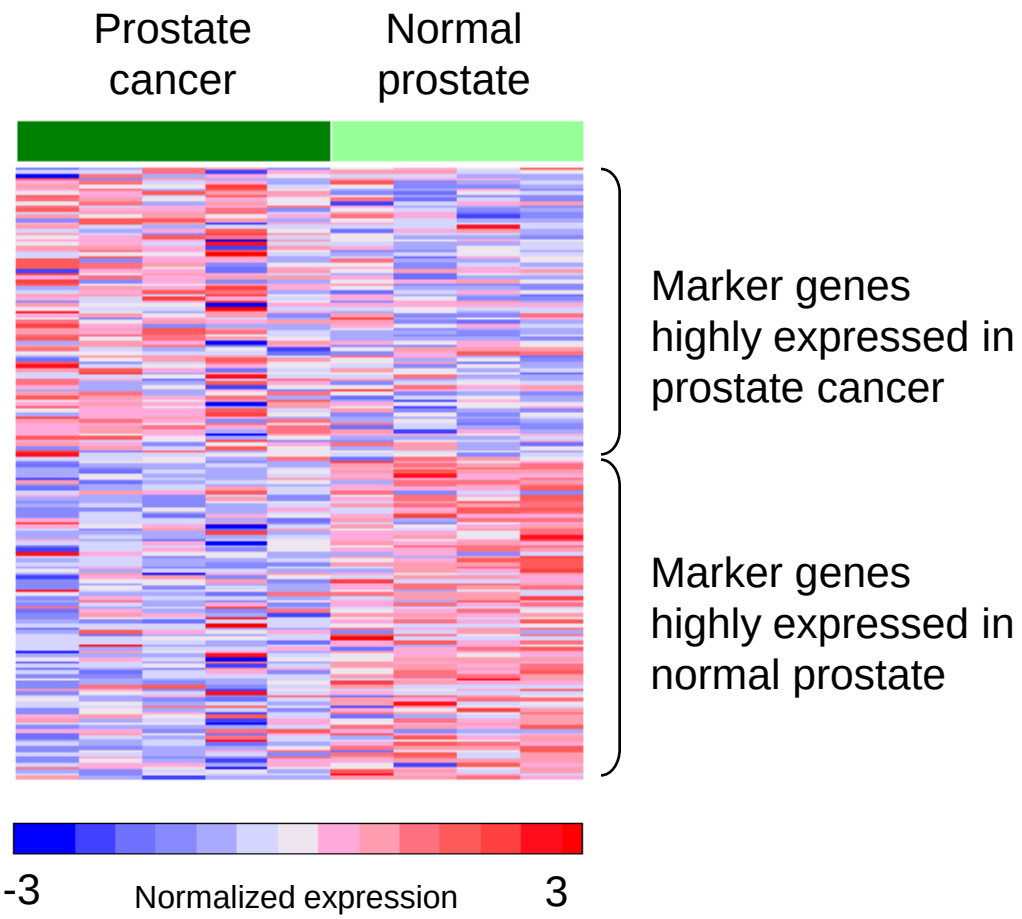
Supplementary Figure 11

(log2 fold change)

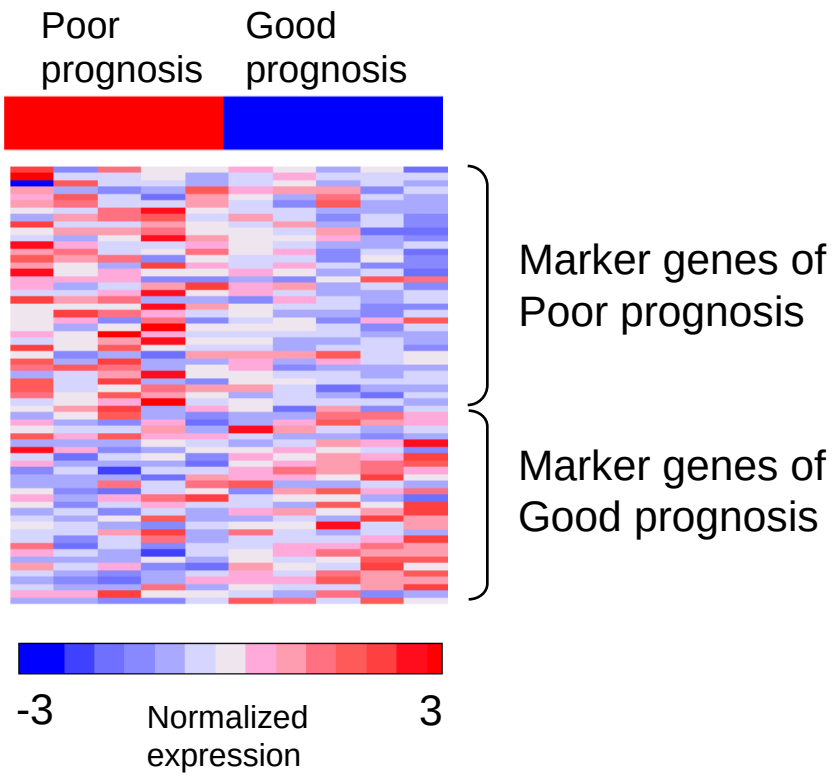


(log2 fold change)

Supplementary Figure 12



Supplementary Figure 13



Supplementary Figure 14